Signal Processing 155 (2019) 170–181

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

The matching pursuit algorithm revisited: A variant for big data and new stopping rules

Fangyao Li [a], Christopher M. Triggs [a], Bogdan Dumitrescu [b], Ciprian Doru Giurcăneanu [a,*]

[a] Department of Statistics, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand
[b] University Politehnica of Bucharest, 313 Spl. Independenței, Bucharest 060042, Romania

ARTICLE INFO

Article history:
Received 22 March 2018
Revised 19 September 2018
Accepted 22 September 2018
Available online 24 September 2018

Keywords:
Matching pursuit algorithm
Hat matrix
Big data
Information theoretic criteria
Air pollution data

ABSTRACT

The matching pursuit algorithm (MPA) is used in many applications for selecting the best predictors for a vector of measurements of size n from a dictionary that contains p_n atoms, where usually $n \leq p_n$. A major unsolved problem is to determine the optimal stopping rule. In this work, we investigate various stopping rules which are modifications of the information theoretic (IT) criteria derived for Gaussian linear regression. Because all of them involve the degrees of freedom (df) given by the trace of the hat matrix, we provide some theoretical results concerning this matrix. We also propose novel stopping rules. An important contribution of this paper is a method for computing the df efficiently when big data ($n \gg p_n$) are processed. The significance of the auxiliary variables appearing in MPA for big data is clarified via a theoretical analysis. The superiority of the new stopping rules in comparison with the traditional approaches is demonstrated in simulations involving big data ($n \gg p_n$) or overcomplete dictionaries ($n < p_n$) and in experiments with air pollution data.

1. Introduction

1.1. Motivation

An important problem in multivariate signal processing is the prediction of a particular entry of the vector random process $\{\mathbf{y}(t)\}$ by using the past measurements as well as the current measurements available for the other entries of the vector (see, for example, [1] and the references therein). The problem can be easily solved by applying the techniques for the identification of autoregressive models with exogenous input [2]. The most difficult part is the selection of the best possible predictors from the existing set of observations. In many practical applications, a large number of past samples are available and this restrains the use of the full-search approach during the training phase when the predictors are chosen.

The computational effort for selecting the predictors can be reduced significantly by applying greedy algorithms [3]. From this family of algorithms, we are especially interested in the matching pursuit algorithm (MPA), which is extensively used in signal processing [4], statistics [5], and approximation theory [6]. At each iteration, MPA yields a linear model for the response vector $\mathbf{y}$ of size n; each such model is a linear combination of some of the entries of a given set of p_n predictors. Theoretical results on the performance of MPA have been recently proven [7] under the hypotheses that (i) p_n grows very fast when n increases and (ii) the predictors are not independent.

The number of iterations for MPA can be as large as $m_{\text{ub}} = 20,000$ and a different model is created at each iteration. The outcome of the algorithm is the model deemed to be "the best" with respect to the selection rule. Because a selection rule decides the outcome of MPA, it is often called the *stopping rule*. An open problem concerns the stopping rule that should be applied as the use of cross-validation (CV) is computationally intensive when the number of iterations, m_{ub}, is large [3].

In our conference paper [8], we have investigated the performance of eleven stopping rules based on different information theoretic (IT) criteria. All of them have been derived from selection rules previously applied in classical linear regression. Another common feature is the presence of the degrees of freedom (df) in their expressions. According to the definition [9], df is evaluated as the trace of the linear operator mapping $\mathbf{y}$ to $\hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ is the estimate of $\mathbf{y}$ produced by a certain model. This linear operator is known as the hat matrix. Importantly [10], there is empirical evidence that the trace-based computation may underestimate the value of df.

* corresponding author.
E-mail addresses: lfan523@aucklanduni.ac.nz (F. Li), cm.triggs@auckland.ac.nz (C.M. Triggs), bogdan.dumitrescu@acse.pub.ro (B. Dumitrescu), c.giurcaneanu@auckland.ac.nz (C.D. Giurcăneanu).

https://doi.org/10.1016/j.sigpro.2018.09.033
0165-1684/

1.2. Contributions

After presenting MPA in Section 2, we outline the following results in the rest of the paper.

In Section 3, we briefly discuss the IT criteria which are currently used in conjunction with MPA and introduce new stopping rules. The new formulae are based on the properties of the hat matrix that are presented in Appendix A. We show that, in general, the hat matrix is not a projector and give an upper bound on the increase of df from the mth iteration of the algorithm to the $(m+1)$th iteration. These results were stated without proof in [8].

It has been already pointed out in [7] that, because of the massive amount of data produced nowadays, a formulation of MPA for $n \gg p_n$ is really needed. Re-writing the algorithm for the big data case is straightforward and it was already done in [7], but the most difficult part is the calculation of df at each iteration. In Theorem 1, we demonstrate how this can be done efficiently. There is no result similar to Theorem 1 in the previous literature. In Section 4 we perform a theoretical analysis that clarifies the significance of the auxiliary variables appearing in the formulation of MPA for big data, but not in the classical formulation of the algorithm.

In Section 5 we present the results of an extensive empirical study which shows the superiority of the newly introduced IT criteria. In experiments with air pollution data, the new criteria work better than CV. A more comprehensive discussion of the theoretical and empirical results obtained in this work can be found in Section 6.

1.3. Notation

Bold letters denote both vectors and matrices; $\mathbf{I}$ denotes the identity matrix of appropriate size, while $\mathbf{0}$ denotes the vector/matrix whose entries are all equal to zero. The symbol x_a stands for the ath entry of a vector $\mathbf{x}$. If $\mathbf{X}$ is a matrix, then $\mathbf{X}_{a:}$ is the ath row of $\mathbf{X}$, $\mathbf{X}_{:b}$ is the bth column of $\mathbf{X}$, and x_{ab} denotes the entry of $\mathbf{X}$ located in the ath row and the bth column. The operator for transposition is $(\cdot)^\top$; the Euclidean norm of a vector $\mathbf{x}$ is $\|\mathbf{x}\|$; the operator $\odot$ is employed for the element-wise product of vectors. For an arbitrary matrix $\mathbf{X}$, $\mathrm{Sp}(\mathbf{X})$ denotes the linear subspace spanned by the columns of $\mathbf{X}$ and $\mathrm{Ker}(\mathbf{X})$ is the null space of $\mathbf{X}$.

2. The matching pursuit algorithm

2.1. Description

Assume that the response vector $\mathbf{y} = [y_1, \ldots, y_n]^\top$ is given, as well as the matrix $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_{p_n}]$ of potential predictors, which is called dictionary. If $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the fitted linear model, then all non-zero entries of $\hat{\boldsymbol{\beta}}$ correspond to the selected predictors. The residuals are given by $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. In the initialization phase of the algorithm, the vector $\mathbf{y}$ and the columns of $\mathbf{X}$ are centred, and $\hat{\boldsymbol{\beta}}$ is set to $\mathbf{0}$. At each iteration, MPA selects the column of $\mathbf{X}$ leading to the largest reduction of the residual sum of squares. Assume that, at the jth step of the algorithm, the column of $\mathbf{X}$ indexed by $s(j)$ is selected, where $1 \le s(j) \le p_n$. Then, only the $s(j)$th entry of $\hat{\boldsymbol{\beta}}$ is updated by using the formula $\hat{\beta}_{s(j)} \leftarrow \hat{\beta}_{s(j)} + \nu(\mathbf{x}_{s(j)}^\top \mathbf{x}_{s(j)})^{-1}\mathbf{x}_{s(j)}^\top \mathbf{e}$. MPA can be seen as a coordinate descent on the objective $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, the chosen coordinate corresponding to the largest element of the gradient.

The parameter $\nu \in (0, 1]$ is the step size, also known as the shrinkage parameter. Note that all other entries of $\hat{\boldsymbol{\beta}}$ remain unchanged. This is a major difference from orthogonal matching pursuit (OMP) which re-estimates all the entries of the vector of linear parameters at each step of the algorithm. The two algorithms have been already compared in [3, Sec. 12.7.1.1].

In general, the value of the shrinkage parameter in MPA is taken to be small, for example, $\nu = 0.1$. This is justified in [3, Sec. 12.6.2.1] by emphasizing the relationship between MPA and the well-known Lasso algorithm [11]. Another peculiarity of MPA is that the same predictor can be selected not only once, but multiple times during the iterations of the algorithm even when $\nu = 1$. This makes it difficult to evaluate the complexity of the linear model produced at each step of MPA. We discuss this aspect below.

2.2. Hat matrix

Let $\hat{\mathbf{y}}_m = \mathbf{X}\hat{\boldsymbol{\beta}}_m$ be the estimate of $\mathbf{y}$ obtained after the mth step of the algorithm. We denote by $\mathbf{B}_m$ the linear operator, named the hat-matrix, which maps $\mathbf{y}$ to $\hat{\mathbf{y}}_m$:

$$\hat{\mathbf{y}}_m = \mathbf{B}_m \mathbf{y}. \tag{1}$$

Recalling that $\mathbf{x}_{s(j)}$ denotes the predictor selected at the jth iteration of MPA, $\mathbf{B}_m$ is expressed as [12] (see also the discussion in [5, Sec. 5.3]):

$$\mathbf{B}_m = \mathbf{I} - \mathbf{A}_m, \text{ where} \tag{2}$$

$$\mathbf{A}_m = \left(\mathbf{I} - \nu\mathbf{P}_{s(m)}\right)\cdots\left(\mathbf{I} - \nu\mathbf{P}_{s(1)}\right), \tag{3}$$

$\mathbf{P}_{s(j)} = \tilde{\mathbf{x}}_{s(j)}\tilde{\mathbf{x}}_{s(j)}^\top$ and $\tilde{\mathbf{x}}_{s(j)} = \mathbf{x}_{s(j)}/\|\mathbf{x}_{s(j)}\|$ for $1 \le j \le m$. It can be shown by mathematical induction that

$$\mathbf{A}_m = \sum_{k=0}^{m} \mathbf{S}_{m,k}, \text{ where } \mathbf{S}_{m,0} = \mathbf{I} \tag{4}$$

and we have for $1 \le k \le m$:

$$\mathbf{S}_{m,k} = (-\nu)^k \sum_{m \ge j_k > j_{k-1} > \cdots > j_1 \ge 1} \mathbf{P}_{s(j_k)}\mathbf{P}_{s(j_{k-1})}\cdots\mathbf{P}_{s(j_1)}. \tag{5}$$

The matrix $\mathbf{B}_m$ is important in evaluating the complexity of the linear model produced at the mth step. More precisely, the degrees of freedom for the fitted model are estimated by

$$\mathrm{df}_m = \mathrm{tr}(\mathbf{B}_m). \tag{6}$$

This formula has been used, for example, in [9]. It follows from Stein's theory on unbiased risk estimation [13] that for the case when the design matrix is fixed and the residuals are i.i.d. Gaussian, with zero-mean and known variance σ^2, $\mathrm{df} = \sum_{j=1}^{n} \mathrm{Cov}(\hat{y}_j, y_j)/\sigma^2$ [14,15]. It is a simple exercise to demonstrate that this expression equals the trace of the hat matrix (see [3, Eq. (2.34)]).

In practice, the user chooses an upper bound, m_{ub}, for the number of iterations. It is often recommended to use an IT criterion for selecting the best model from the m_{ub} different models produced during these iterations. Because of the particularities of MPA, the IT criteria that have been previously derived for the classical linear model cannot be applied in their original form [12]. The modifications of the criteria are discussed in Section 3. They are based on the properties of the hat matrix outlined in Appendix A.

3. Modified IT criteria

We consider the classical linear regression problem for which the additive noise is i.i.d. zero-mean Gaussian, with unknown variance. Let $\hat{\boldsymbol{\beta}}_\gamma$ denote the estimated vector of linear parameters for a model whose set of regressor variables is γ. The vector of residuals is $\mathbf{e}_\gamma = \mathbf{y} - \hat{\mathbf{y}}_\gamma$, where $\hat{\mathbf{y}}_\gamma$ is the estimate calculated by using $\hat{\boldsymbol{\beta}}_\gamma$. We denote the cardinality of γ by $|\gamma|$, and assume that $|\gamma| > 0$. This means that we exclude the possibility that $\mathbf{y}$ is pure noise. An

Table 1
IT criteria: formulae for the classical linear regression problem with $|\gamma|>0$ and the references where they have been derived. In order to make them compatible with MPA, the following alterations are applied to all criteria: $||\mathbf{e}_\gamma||^2 \mapsto ||\mathbf{e}_m||^2$ and $|\gamma| \mapsto \mathrm{df}_m$. Modifications applied only to some of the criteria are listed in the second column. The modified criteria are given in the third column by mentioning the works from where they have been taken. Note that $\psi(\cdot)$ stands for the digamma function.

Original criterion	Alterations		Modified criterion
	$\|\|\hat{\mathbf{y}}_\gamma\|\|^2 \downarrow \|\|\hat{\mathbf{y}}_m\|\|^2$	$\|\|\hat{\mathbf{y}}_\gamma\|\|^2 \downarrow \|\|\mathbf{y}\|\|^2 - \|\|\mathbf{e}_m\|\|^2$	
Akaike Information Criterion (corrected) [23,24] $\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n} + \frac{1+\|\gamma\|/n}{1-(\|\gamma\|+2)/n}$			AIC_C [12]
Kullback Information Criterion [25] $n\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n} + 3\|\gamma\|$			KIC
Kullback Information Criterion (corrected) [26] $n\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n} + \frac{2(\|\gamma\|+1)n}{n-\|\gamma\|-2} - n\psi(\frac{n-\|\gamma\|}{2})$			KIC_C
Bayesian Information Criterion [27] $n\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n} + \|\gamma\|\ln n$			BIC [3]
Stochastic Complexity (SC) [28]	✓		SC_1 [8]
		✓	SC_2 [8]
$n\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n} + \|\gamma\|\ln\frac{\|\|\hat{\mathbf{y}}_\gamma\|\|^2/|\gamma|}{\|\|\mathbf{e}_\gamma\|\|^2/(n-|\gamma|)} + \ln\frac{|\gamma|}{(n-|\gamma|)^{n-1}}$			
Generalized Minimum Description	✓		gMDL_1 [8]
Length [29]		✓	gMDL_2 [5]
If $\frac{\|\|\hat{\mathbf{y}}_\gamma\|\|^2}{\|\|\mathbf{y}\|\|^2} < \frac{|\gamma|}{n}$, $\frac{n}{2}\ln\frac{\|\|\mathbf{y}\|\|^2}{n} + \frac{1}{2}\ln n$, Else $\frac{n}{2}\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n-|\gamma|} + \frac{|\gamma|}{2}\ln\frac{\|\|\hat{\mathbf{y}}_\gamma\|\|^2/|\gamma|}{\|\|\mathbf{e}_\gamma\|\|^2/(n-|\gamma|)} + \ln n$			
Minimum Message Length [30] (uniform prior) $\frac{n-\|\gamma\|}{2}\ln(2\pi) + \frac{n-\|\gamma\|}{2}(1+\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n-\|\gamma\|}) + \frac{\|\gamma\|}{2}\ln(\pi\|\|\mathbf{y}\|\|^2) - \ln\Gamma(1+\frac{\|\gamma\|}{2}) + \frac{1}{2}\ln(\|\gamma\|+1)$			MMLU
Minimum Message Length [30]	✓		MMLG_1
(g-prior)		✓	MMLG_2
If $\frac{\|\|\hat{\mathbf{y}}_\gamma\|\|^2}{\max(\|\gamma\|-2,1)} \le \frac{\|\|\mathbf{e}_\gamma\|\|^2}{n-\|\gamma\|+2}$, $\frac{n}{2}(1+\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n-\|\gamma\|+2}) + \frac{1}{2}\ln(n-1) + \frac{1}{2}$, Else $\frac{n-\|\gamma\|+2}{2}(1+\ln\frac{\|\|\mathbf{e}_\gamma\|\|^2}{n-\|\gamma\|+2}) + \frac{\|\gamma\|-2}{2}\ln(\frac{\|\|\hat{\mathbf{y}}_\gamma\|\|^2}{\max(\|\gamma\|-2,1)}) + \frac{1}{2}\ln((n-\|\gamma\|)\|\gamma\|^2)$			

IT criterion is evaluated for all models of interest, and the model which minimizes the criterion is deemed to be the "best".

It has been already mentioned in the previous literature that these criteria cannot be applied straightforwardly for selecting the number of iterations in MPA. In most cases, the criteria are altered as follows: $||\mathbf{e}_\gamma||^2 \mapsto ||\mathbf{e}_m||^2$ and $|\gamma| \mapsto \mathrm{df}_m$, where the entries of $\mathbf{e}_m$ are the residuals computed at the mth step of the MPA. We note that $\mathrm{df}_m \ge 0$: According to Result 2 in Appendix A, the magnitudes of the eigenvalues of $\mathbf{A}_m$ are not larger than one. It follows from (2) that the hat matrix satisfies $\mathrm{tr}(\mathbf{B}_m) \ge 0$.

A more complicated approach takes into consideration Remark 4 in Appendix A and replaces $||\hat{\mathbf{y}}_\gamma||^2$ either with $||\hat{\mathbf{y}}_m||^2$ or with $||\mathbf{y}||^2 - ||\mathbf{e}_m||^2$. In Table 1, we give the formulae of eight important IT criteria and their modified variants.

According to Refs. [16,17], for classical linear regression when the total number of models is large, a supplementary term should be added to the IT criteria. Based on this, the following "extended" criteria were introduced in [8]:

$$\mathrm{EBIC}(\mathbf{y}; m) = \mathrm{BIC}(\mathbf{y}; m) + 2\nu\ln\vartheta, \tag{7}$$

$$\mathrm{ESC}_{\mathrm{alt}}(\mathbf{y}; m) = \mathrm{SC}_{\mathrm{alt}}(\mathbf{y}; m) + 2\nu\ln\vartheta, \tag{8}$$

$$\mathrm{EgMDL}_{\mathrm{alt}}(\mathbf{y}; m) = \mathrm{gMDL}_{\mathrm{alt}}(\mathbf{y}; m) + \nu\ln\vartheta, \tag{9}$$

where $\mathrm{alt} = 1, 2$ and $\ln(\cdot)$ denotes the natural logarithm. The expressions of BIC, $\mathrm{SC}_{\mathrm{alt}}$ and $\mathrm{gMDL}_{\mathrm{alt}}$ can be found in Table 1. Note that

$$\vartheta = \binom{p_n}{s_m} = \frac{\Gamma(p_n+1)}{\Gamma(s_m+1)\Gamma(p_n-s_m+1)}, \tag{10}$$

where s_m is the number of non-zero entries of $\hat{\boldsymbol{\beta}}_m$ and $\Gamma(\cdot)$ denotes the Gamma function.

The presence of the ν-factor in (7) is justified by Proposition 2 in Appendix A which says that the difference between the penalty terms of $\mathrm{BIC}(\mathbf{y}; m+1)$ and $\mathrm{BIC}(\mathbf{y}; m)$ is at most $\nu\ln n$. This suggests that the "additional" penalty term of EBIC should be multiplied by ν. Taking into consideration the well-known relationship between the penalties of SC, gMDL and the penalty of BIC (see, for example [18]), we obtain the expressions of the criteria in (8) and (9).

Another modification of the "extended" criteria was suggested by one of the reviewers and consists in replacing s_m with df_m in (10). The novel criteria are:

$$\mathrm{EBIC}^*(\mathbf{y}; m) = \mathrm{BIC}(\mathbf{y}; m) + 2\ln\vartheta^*, \tag{11}$$

$$\mathrm{ESC}^*_{\mathrm{alt}}(\mathbf{y}; m) = \mathrm{SC}_{\mathrm{alt}}(\mathbf{y}; m) + 2\ln\vartheta^*, \tag{12}$$

$$\mathrm{EgMDL}^*_{\mathrm{alt}}(\mathbf{y}; m) = \mathrm{gMDL}_{\mathrm{alt}}(\mathbf{y}; m) + \ln\vartheta^*, \tag{13}$$

where $\mathrm{alt} = 1, 2$ and

$$\vartheta^* = \frac{\Gamma(p_n+1)}{\Gamma(\mathrm{df}_m+1)\Gamma(p_n-\mathrm{df}_m+1)}. \tag{14}$$

We also introduce a new criterion, inspired from a previous work on graphical models. More precisely, we modify the criterion in [19, Eq. (1)] by using df_m instead of the number of edges, and by replacing the number of nodes of the graphical model with p_n. After these alterations, we get the following formula:

$$\mathrm{EBIC}^\circ(\mathbf{y}; m) = \mathrm{BIC}(\mathbf{y}; m) + 4\mathrm{df}_m\ln p_n. \tag{15}$$

The expression above can be regarded as a modified variant of the criterion proposed in [20, Eq. (2)] for model selection in linear regression when the number of measurements is much smaller than

the total number of predictors. For obtaining (15), we take the constant c in [20, Eq. (2)] to be two and apply the transformations $||\mathbf{e}_\gamma||^2 \mapsto ||\mathbf{e}_m||^2$, $|\gamma| \mapsto \mathrm{df}_m$. More importantly, instead of the determinant of the Fisher information matrix, we use its asymptotic approximation. This improves the performance of model selection in the presence of collinearity [21], but has a negative effect when the variance of the additive noise tends to zero. As the noiseless case is of little interest for this work, we only mention here that both SC and gMDL can handle it without difficulty (see [18]).

For understanding the relationship between EBIC* and EBIC°, we compare the penalty terms of the two criteria when $p_n \gg \mathrm{df}_m$. To this end, we apply the Stirling approximation $\ln\Gamma(z) = \left(z-\frac{1}{2}\right)\ln z - z + \frac{1}{2}\ln(2\pi)$ [22] to $\Gamma(p_n+1)$ and $\Gamma(p_n-\mathrm{df}_m+1)$ in (14). Hence, the penalty term of EBIC* can be written as $2\ln\vartheta^* = 2\mathrm{df}_m \ln(p_n-\mathrm{df}_m+1) + o(\mathrm{df}_m \ln p_n)$, which implies that the penalty term of EBIC* is half of the penalty term of EBIC° when $p_n \gg \mathrm{df}_m$.

In comparison with EBIC, EBIC° has the advantage that its penalty term does not decrease when $s_m > p_n/2$ (see [8]). It is worth mentioning that EBIC° reduces to BIC when $n \gg p_n^4$. This is part of the big data case, for which we provide below a computationally efficient approach.

4. MPA for big data

4.1. Modified algorithm

A formulation of MPA for big data is proposed in [7], where the algorithm is written for the case when $n \gg p_n$. The key point is to keep in memory only the vector $\mathbf{c} = (\mathbf{X}^\top\mathbf{y})/n$ of length p_n and the matrix $\mathbf{D} = (\mathbf{X}^\top\mathbf{X})/n$ of size $p_n \times p_n$. Hence, all the calculations involved by the algorithm should be done by only using the entries of $\mathbf{c}$ and $\mathbf{D}$, and without having access to the entries of $\mathbf{X}$ and $\mathbf{y}$.

For better understanding how this can be done, we analyse in detail all the calculations performed at the $(m+1)$th step of the algorithm, where $0 \le m < m_{\mathrm{ub}}$. We assume that the vector $\mathbf{y}$ and the columns of $\mathbf{X}$ are centred; the columns of $\mathbf{X}$ are standardised such that all the diagonal entries of $(\mathbf{X}^\top\mathbf{X})/n$ are equal to one. Moreover, the algorithm is initialized as follows: $m \leftarrow 0$ and $\hat{\boldsymbol{\beta}} \leftarrow \mathbf{0}$.

Let $\hat{\boldsymbol{\beta}}_m$ be the vector of linear parameters estimated at the mth step; the corresponding residuals are the entries of $\mathbf{e}_m = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_m$. For any index j $(1 \le j \le p_n)$, the squared norm of the vector of residuals produced by selecting the jth predictor is $||\mathbf{e}_m||^2 - (\mathbf{x}_j^\top\mathbf{e}_m)^2/n$. As we are interested in finding the predictor that leads to the largest reduction of the sum of squares, we take $s(m+1) = \arg\max_{1\le j\le p_n} |\mathbf{x}_j^\top\mathbf{e}_m|$. Let $\boldsymbol{\alpha}_m = \mathbf{c} - \mathbf{D}\hat{\boldsymbol{\beta}}_m = (\mathbf{X}^\top\mathbf{e}_m)/n$. The index $s(m+1)$ corresponds to the entry of $\boldsymbol{\alpha}_m$ which has the largest magnitude. All the entries of $\hat{\boldsymbol{\beta}}_{m+1}$ are the same as the entries of $\hat{\boldsymbol{\beta}}_m$, except one which is updated as follows: $\hat{\beta}_{s(m+1)} \leftarrow \hat{\beta}_{s(m+1)} + \nu\alpha_{m,s(m+1)}$. Note that $\alpha_{m,s(m+1)}$ is the $s(m+1)$th entry of the vector $\boldsymbol{\alpha}_m$. This is the procedure used in [7] for selecting the predictor and updating the $\hat{\boldsymbol{\beta}}$-vector.

However, for the evaluation of the IT criteria we need some quantities that are not computed in [7]: $||\hat{\mathbf{y}}_{m+1}||^2/n$ is readily obtained and then can be used in the following calculations: $||\mathbf{e}_{m+1}||^2/n = (||\mathbf{y}||^2 + ||\hat{\mathbf{y}}_{m+1}||^2)/n - 2\hat{\boldsymbol{\beta}}_{m+1}^\top\mathbf{c}$. It is easy to observe that MPA can be implemented such that the computational complexity for each iteration is $\mathcal{O}(p_n)$.

4.2. Computation of the degrees of freedom

We need to evaluate $\mathrm{df}_{m+1} = \mathrm{tr}(\mathbf{B}_{m+1}) = n - \mathrm{tr}(\mathbf{A}_{m+1})$. Although we have from (3) that

$$\mathbf{A}_{m+1} = \left(\mathbf{I} - \nu\mathbf{P}_{s(m+1)}\right)\mathbf{A}_m, \tag{16}$$

it is not straightforward to calculate df_{m+1} when the matrix $\mathbf{X}$ is not stored in the memory. Even if $\mathbf{X}$ were available, the explicit computation of $\mathbf{A}_{m+1}$ requires $\mathcal{O}(n^2)$ operations. A computationally efficient solution is presented in the theorem below. To this end, we need to introduce the auxiliary variables $\mathbf{v}, \mathbf{w} \in \mathbb{R}^{p_n}$ and $\mathbf{G} \in \mathbb{R}^{p_n \times p_n}$. MPA is initialized such that df is zero, and all the entries of $\mathbf{w}$ and $\mathbf{G}$ are equal to zero.

Theorem 1. *The degrees of freedom for the model produced at the $(m+1)$ th step of the MPA $(0 \le m < m_{\mathrm{ub}})$ can be computed by using the updating formulae:*

$$\mathbf{v}^\top \leftarrow -\nu\mathbf{D}_{s(m+1):}\mathbf{G} + \nu^2\mathbf{D}_{s(m+1):} \odot \mathbf{w}^\top, \tag{17}$$

$$\mathbf{G}_{s(m+1):} \leftarrow \mathbf{G}_{s(m+1):} + \mathbf{v}^\top, \tag{18}$$

$$w_{s(m+1)} \leftarrow w_{s(m+1)} + 1, \tag{19}$$

$$\mathrm{df}_{m+1} \leftarrow \mathrm{df}_m + \nu - \mathbf{v}^\top\mathbf{D}_{:s(m+1)}. \tag{20}$$

The proof is deferred to Appendix B.

One can easily see that the number of operations for computing df_{m+1} decreases from $\mathcal{O}(n^2)$ to $\mathcal{O}(p_n^2)$ if Eqs. (17)–(20) are employed instead of applying (16). As $n \gg p_n$, it means that the recursions in Theorem 1 allow reduction of the memory usage and, at the same time, improve the execution time.

It follows from Theorem 1 that the entries of $\mathbf{w}$ are counts for how many times each predictor is selected. However, the significance of $\mathbf{v}$ and $\mathbf{G}$ is not very clear. To gain more insight, we prove the following results in Appendix B.

4.3. Some properties of the variables $\mathbf{v}$ and $\mathbf{G}$

Given a non-negative integer m_0, we denote by $\ell(1), \ldots, \ell(\mu)$ the iterations of MPA at which the first predictor $(\tilde{\mathbf{x}}_1)$ is selected, with $m_0 < \ell(1) < \cdots < \ell(\mu) \le m_{\mathrm{ub}}$. For ease of writing, we take $\mathbf{d}^\top = [1\ \tilde{\mathbf{d}}^\top]$ to be the first row of $\mathbf{D}$. Similarly, $\mathbf{g}_{\ell(j)}^\top$ is the first row of the $\mathbf{G}$-matrix obtained at the $\ell(j)$th iteration such that $\mathbf{G}_{\ell(j)} = [\mathbf{g}_{\ell(j)}\ \tilde{\mathbf{G}}_{\ell(j)}^\top]^\top$. The symbol $\mathbf{v}_{\ell(j)}$ stands for the vector $\mathbf{v}$ evaluated at the $\ell(j)$th iteration by using (17). Similarly, $\mathbf{w}_{\ell(j)}$ denotes the vector $\mathbf{w}$ evaluated at the $\ell(j)$th iteration by applying (19).

Lemma 1. *(i) For $1 \le j \le \mu - 1$, the following identities hold:*

$$\mathbf{g}_{\ell(j+1)}^\top = (1-\nu)\mathbf{g}_{\ell(j)}^\top + \nu^2(j+j_0)\mathbf{h}^\top + \mathbf{r}_{\ell(j+1)}^\top, \tag{21}$$

$$\mathbf{v}_{\ell(j+1)}^\top = (1-\nu)\mathbf{v}_{\ell(j)}^\top + \nu^2\mathbf{h}^\top + \mathbf{r}_{\ell(j+1)}^\top - \mathbf{r}_{\ell(j)}^\top, \tag{22}$$

where j_0 represents how many times the predictor $\mathbf{x}_1$ was selected during the first m_0 iterations, and $\mathbf{h}^\top$ is the first row of the identity matrix of size $p_n \times p_n$. We define $\mathbf{r}_{\ell(j)}^\top = -\nu\tilde{\mathbf{d}}^\top\tilde{\mathbf{G}}_{\ell(j)-1} + \nu^2[0\ \tilde{\mathbf{d}}^\top \odot \tilde{\mathbf{w}}_{\ell(j)-1}^\top]$, where $\tilde{\mathbf{w}}_{\ell(j)-1}^\top$ contains the last $p_n - 1$ entries of $\mathbf{w}_{\ell(j)-1}^\top$.

(ii) From (22), we get:

$$\mathbf{v}_{\ell(\mu)}^\top = (1-\nu)^{\mu-1}\mathbf{v}_{\ell(1)}^\top + \nu[1-(1-\nu)^{\mu-1}]\mathbf{h}^\top + \sum_{j=2}^{\mu}(\mathbf{r}_{\ell(j)}^\top - \mathbf{r}_{\ell(j-1)}^\top)(1-\nu)^{\mu-j}. \tag{23}$$

These findings are instrumental in proving the next proposition.

Proposition 1. *If* $\lim_{\mu\to\infty}(\mathbf{r}_{\ell(\mu)}^{\top}-\mathbf{r}_{\ell(\mu-1)}^{\top})=\mathbf{0}$, *then we have:*

$$\lim_{\mu\to\infty}\mathbf{v}_{\ell(\mu)}^{\top}=\nu\mathbf{h}^{\top}, \tag{24}$$

$$\lim_{\mu\to\infty}\frac{\mathbf{g}_{\ell(\mu)}^{\top}}{\mu}=\nu\mathbf{h}^{\top}. \tag{25}$$

The condition which appears in Proposition 1 is clearly satisfied whenever the first predictor is the only predictor selected in a long sequence of iterations.

4.4. Analysis of the case when m is large

The condition in Proposition 1 is also satisfied when m_0 is large. It is well-known that the vector of linear parameters estimated at the mth step, $\hat{\boldsymbol{\beta}}_m$, converges to the least-squares solution $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ when m tends to infinity [5, Sec. 4.1]. Therefore, for large μ, $\mathbf{r}_{\ell(\mu)}^{\top}$ is almost the same as $\mathbf{r}_{\ell(\mu-1)}^{\top}$, and this property implies that $\mathbf{v}_{\ell(\mu)}\approx\nu\mathbf{h}^{\top}$ and $\mathbf{g}_{\ell(\mu)}^{\top}/\mu\approx\nu\mathbf{h}^{\top}$.

The reasoning applied to the first predictor can be extended to all other predictors. For large m, we have:

- If the jth predictor is selected at the mth iteration, then $\mathbf{v}$ is approximately equal to the jth row of $\nu\mathbf{I}$.
- If at the mth iteration we divide each entry of the jth row of $\mathbf{G}$ by the jth entry of $\mathbf{w}$ for all $j\in\{1,\dots,p_n\}$, then the resulting matrix is approximately equal to $\nu\mathbf{I}$.
- As $\hat{\boldsymbol{\beta}}_m\approx\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$, we get from (1) that $\mathbf{B}_m$ is approximately equal to the orthogonal projector onto $\mathrm{Sp}\big(\tilde{\mathbf{x}}_1,\dots,\tilde{\mathbf{x}}_{p_n}\big)$. It follows from (6) and [31, Lemma 2.2] that $\mathrm{df}_m\approx\mathrm{rank}(\mathbf{X})$. We have from (20) that $\mathrm{df}_{m+1}\approx\mathrm{df}_m$ because of (24).

After these theoretical considerations, we conduct an empirical study for evaluating the performance of MPA.

5. Experimental results

5.1. Simulated data

The procedure for simulating the data is similar to the one in [7, Sec. 3.3], and comprises the following steps.

5.1.1. Generation of the dictionary $\boldsymbol{X}$ of size $n\times p_n$

For an arbitrary $T>n$, let $\mathbf{E}$ be a matrix with $T+n$ rows. The columns of $\mathbf{E}$ are realizations of the autoregressive (AR) process $\mathbf{E}_{:j}=\omega\mathbf{E}_{:j-1}+\mathbf{u}_j$, where $\omega\in(-1,1)$ and $j\in\{1,\dots,p_n\}$. It is obvious that the model represents a proper AR process only when $\omega\neq0$. Additionally, the random vectors $\{\mathbf{u}_j\}$ are i.i.d. Gaussian, with mean $\mathbf{0}$ and covariance matrix $(1-\omega^2)\mathbf{I}$. In order to reduce the effect of initialization, we generate a sequence of $(p_n+100)\mathbf{E}_{:j}$-vectors and keep only the last p_n in $\mathbf{E}$. These are used for producing the rows of $\mathbf{X}$:

$$\mathbf{X}_{i:}=\sum_{t=0}^{T}\theta_t\mathbf{E}_{i+T-t:},\ \text{for } i=\overline{1,n}. \tag{26}$$

The coefficients θ_t are such that $1=\theta_0\geq\theta_1\geq\cdots\geq\theta_T\geq0$.

In our settings, $\omega=0$ or $\omega=0.75$. The following cases are considered: *Case I* (memoryless): $\theta_t=0$ if $t>0$; *Case II* (short memory, geometric decay): $T=100+n$ and $\theta_t=0.95^t$, for $0<t\leq T$; *Case III* (long memory, slow decay): $T=1000+n$ and $\theta_t=(t+1)^{-1/2}$, for $0<t\leq T$.

5.1.2. Generation of the response vector $\boldsymbol{y}$ of length n

It is given by $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, where the entries of $\boldsymbol{\beta}$ are chosen according to four different models: *Model 1* (low-dimensional): $\beta_1=\beta_2=\beta_3=1/3$ and $\beta_q=0$ for $4\leq q\leq p_n$; *Model 2* (high-dimensional, small equal coefficients): $\beta_q=p_n^{-1}$ for $1\leq q\leq p_n$; *Model 3* (high-dimensional, decaying coefficients): $\beta_q=q^{-1}$ for $1\leq q\leq p_n$; *Model 4* (high-dimensional, slowly decaying coefficients): $\beta_q=q^{-1/2}$ for $1\leq q\leq p_n$. Remark that only *Model 1* is sparse.

The $\boldsymbol{\varepsilon}$-vector is simulated as follows: Let $\tilde{\boldsymbol{\varepsilon}}$ be a vector of length n obtained by applying the linear filter with coefficients $\{\theta_t\}_{t=0}^{T}$ from (26) to a vector of length $T+n$ whose entries are i.i.d. standard Gaussian random variables. The entries of $\tilde{\boldsymbol{\varepsilon}}$ are statistically independent with respect to the entries of the matrix $\mathbf{E}$. With the convention that $\kappa=\left[\frac{\mathrm{Var}(\mathbf{X}\boldsymbol{\beta})}{\mathrm{Var}(\tilde{\boldsymbol{\varepsilon}})}\right]^{1/2}$ and ς is a parameter that controls the signal-to-noise ratio (SNR), we have: $\boldsymbol{\varepsilon}=(\kappa/\varsigma)\tilde{\boldsymbol{\varepsilon}}$. Following [7, Sec. 3.3], we take $\varsigma^2=8$ for high SNR and $\varsigma^2=0.2$ for low SNR.

5.1.3. Other details of the implementation

The vector $\mathbf{y}$ and the columns of $\mathbf{X}$ are centred. Additionally, the columns of $\mathbf{X}$ are standardised such that all the diagonal entries of $(\mathbf{X}^{\top}\mathbf{X})/n$ are equal to one. The upper bound on the number of iterations for MPA is $m_{\mathrm{ub}}=20,000$. Because of the way in which the expression of AIC_C depends on df (see Table 1), we end the iterations before df equals $n-2$. An additional rule is applied such that MPA is stopped after the number of distinct selected predictors becomes equal to p_n. In our simulation study, we take $\nu=0.1$ and this choice is based on the findings from [3, Sec. 12.6.2.1] [8].

5.1.4. Performance evaluation

In order to test the predictive power of each IT criterion, we use the same method as in [7, Sec. 3.3]: For each trial, the same algorithm as the one used to generate the dictionary is applied in order to produce a matrix $\mathbf{X}_{\mathrm{out},\,r}$ whose size is $(10n)\times p_n$. If $\hat{\boldsymbol{\beta}}_r^{\mathrm{ITC}}$ is the vector of linear parameters corresponding to the model selected by a particular IT criterion, in trial r, then we compute the mean integrated squared error as follows [7, Sec. 3.3]:

$$\mathrm{MISE}=\frac{\sum_{r=1}^{\mathrm{N}_{TR}}\left\|\mathbf{X}_{\mathrm{out},r}\boldsymbol{\beta}-\mathbf{X}_{\mathrm{out},r}\hat{\boldsymbol{\beta}}_r^{\mathrm{ITC}}\right\|^2}{(10n)\times\mathrm{N}_{TR}}, \tag{27}$$

where $\boldsymbol{\beta}$ is defined for each model in the description above and N_{TR} denotes the number of trials. Note that, for all $1\leq r\leq\mathrm{N}_{TR}$, the columns of $\mathbf{X}_{\mathrm{out},\,r}$ are centred.

We fix $p_n=100$ and vary the sample size such that $n\in\{20,100,10,000\}$. For each quintuple (Case, Model, ω, ς^2, n), we compute MISE for each IT criterion from $\mathrm{N}_{TR}=100$ trials. The results are reported in the supplemental material [32]. As we are especially interested in comparing the performance of various criteria, we apply the following scoring: For a fixed (Case, Model, ω, ς^2, n), the criterion which produces the minimum MISE gets one point, any other criterion whose MISE is within 5% from the minimum MISE gets one half of a point, and all other criteria get zero points. When the number of points earned in various experiments are aggregated, the final score is computed as the ratio of total points to the number of experiments. This guarantees that the maximum possible value for the final score is one, and this is earned only by a criterion ranked the best in each of the experiments.

The number of experiments which are done for a fixed sample size is $3\times4\times2\times2=48$. The number of trials for each experiment is $\mathrm{N}_{TR}=100$. We compute the aggregated scores from the 48 experiments conducted for $n=20$ and plot them in Fig. 1. In the

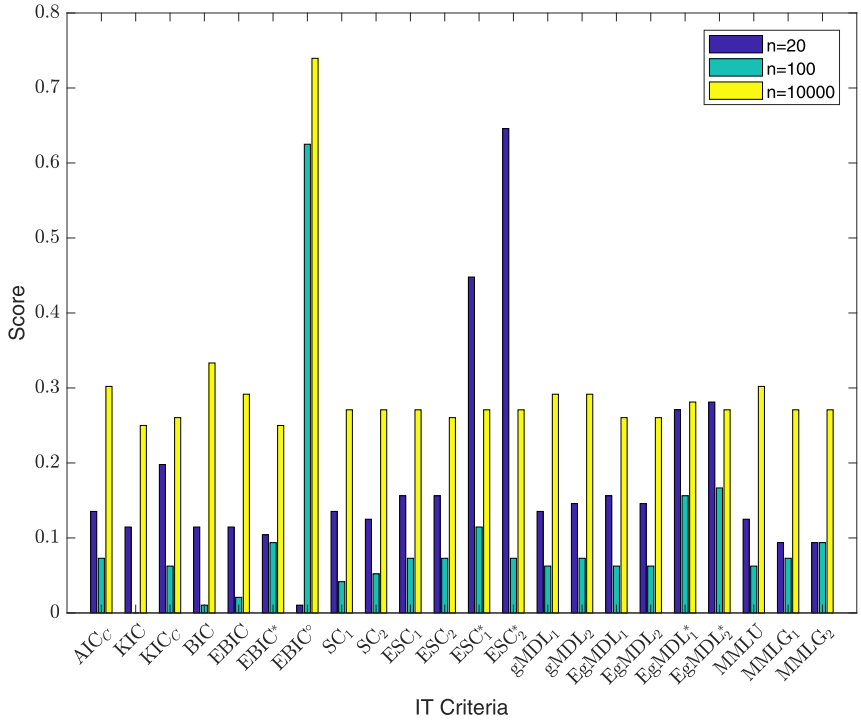

Fig. 1. Scores aggregated from all experiments. Note that the scores are normalized in order to take values in the interval [0,1]. The greater the score, the better the performance of the IT criterion.

same figure, we present the scores which are similarly calculated for $n = 100$ and $n = 10,000$, respectively. It is clear that the best scores are obtained by some of the "extended" criteria: ESC_1^* and ESC_2^* perform well when $n < p_n$, whereas EBIC° is very good when $n \geq p_n$. It is interesting to observe that the criteria in (11)–(13) for which the extra-penalty term is computed by using ϑ^* are superior to those in (7)–(9), where the extra-penalty term involves ϑ.

According to the results reported in [32], the best stopping rules for $n = 20$ are those which select the smallest number of predictors. Based on the analysis that employs Stirling approximation (see Section 3), we expect EBIC° to choose fewer atoms than EBIC*. The experimental results confirm this, especially in the Cases 2 and 3 when $\omega = 0.75$ and $\varsigma^2 = 0.2$. However, the penalty term that depends explicitly on the data makes the difference between EBIC* and $\text{ESC}_{\text{alt}}^*$ (alt=1, 2): because of this term, ESC_1^* and ESC_2^* select models with higher sparsity than those chosen by the other criteria and this explains their good performance. Observe that EgMDL_1^* and EgMDL_2^* produce moderately good results when $n = 20$. This is a supplementary confirmation that the use of the penalty term which involves $\ln \vartheta^*$ has beneficial effects when $p_n > n$. One can further exploit this idea by adding such penalties to MMLU, MMLG_1 and MMLG_2; this is perfectly justified by [30, Sec. 5]. The fact that the number of criteria in our comparison is very large discouraged us from considering this approach. Without the additional penalty term, MMLU, MMLG_1 and MMLG_2 behave like SC_1 and SC_2.

For better understanding how the stopping rules work in specific conditions, we aggregate the scores for a particular Case, or for a certain level of correlation between the columns of the dictionary, or for a particular SNR. In Fig. 2, we show the scores aggregated from all the experiments in which SNR is low ($\varsigma^2 = 0.2$). Note that the best criteria are the same as in Fig. 1, but this time the gap between them and the rest of the stopping rules is much bigger. When comparing the scores for Case 1 that are presented in Fig. 3, ESC_2^* is ranked first when $n = 20$; this is similar to the results shown in Figs. 1 and 2. The main difference occurs in the ranking of EBIC°, which is not the best criterion for $n > 20$.

5.1.5. Comparison with cross-validation (CV)

The implementation is the same leave-one-out cross-validation as in [7, Sec. 3.3]: We sample without replacement $n - 1$ measurements. Then MPA is applied to these measurements and, at each iteration of the algorithm, the estimated linear parameters are used to compute an estimate for the nth measurement. The squared error between this estimate and the "true" value of the nth measurement is calculated. The procedure is repeated $\rho = 25$ times, and the squared errors computed at each iteration are averaged. The selected model is the one which corresponds to the minimum average. Furthermore, MISE is evaluated by applying the same methodology as in the case of IT criteria.

The performance of CV is compared to that of IT criteria in Fig. 4. More precisely, for a given context, we pick the IT criterion, which was found to be the best. For example, in Case 1 we have from Fig. 3 that ESC_2^* is ranked first when $n = 20$. Hence, for each experiment done in Case 1 when $n = 20$, we compare the MISE calculated for ESC_2^* with the one computed for CV. The method with the smallest MISE gets one point. Any other method gets one half of a point if its MISE is within 5% from the MISE of the winner, or zero points otherwise. The final scores are aggregated as it was already explained above.

Based on the results presented in Fig. 4, we can conclude that the performance of CV tends to be superior to that of the IT criteria. The superiority of CV is less evident for $n = 20$, where ESC_2^* is better than CV in the following contexts: overall, high correlation between the columns of the dictionary ($\omega = 0.75$), low SNR ($\varsigma^2 = 0.2$), Case 1 and Case 3. For $n > 20$, there are two contexts in which the IT criteria are better than CV: (i) $n = 100$–EBIC° has a higher score than CV when SNR is low; (ii) $n = 10,000$–AIC_C is superior to CV in Case 1. For fairness, we should mention that the computational complexity of CV is about $\rho = 25$ times higher than the computational complexity of any IT criterion.

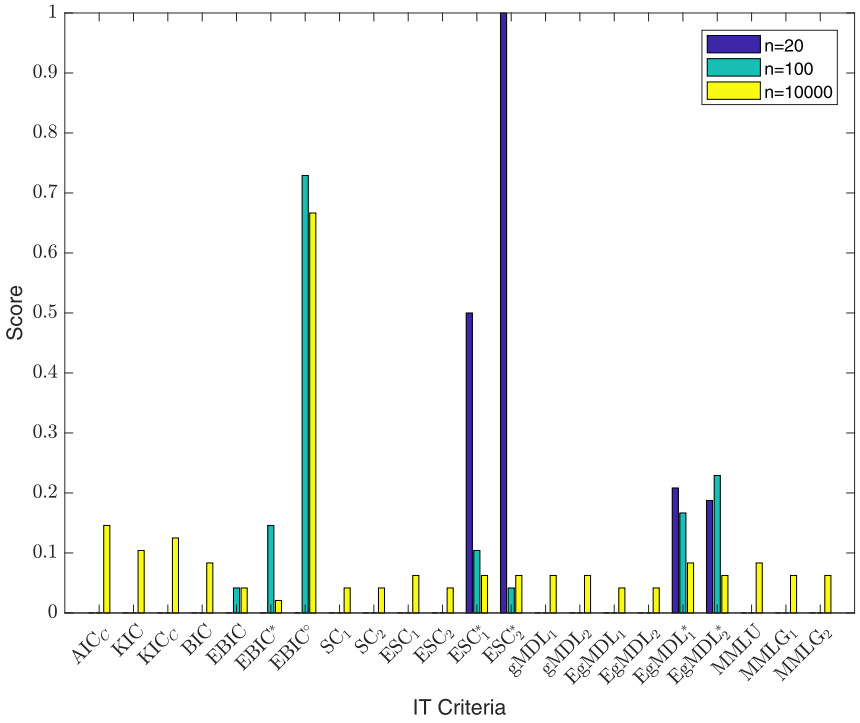

Fig. 2. Scores aggregated from all experiments in which SNR is low ($\varsigma^2 = 0.2$).

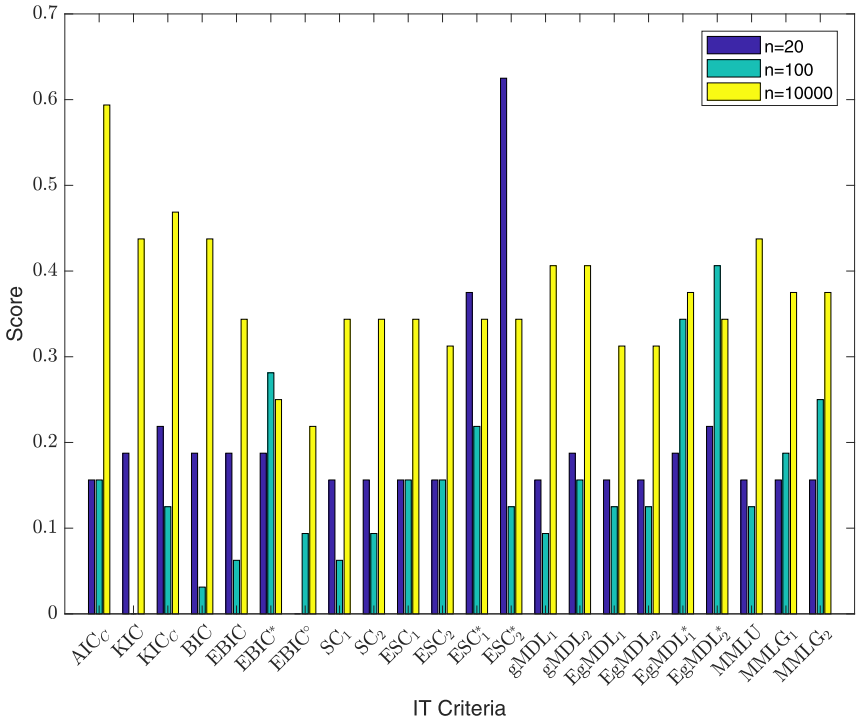

Fig. 3. Scores aggregated from all experiments done for Case 1.

5.2. Air pollution data

5.2.1. Problem formulation

Various studies published in the medical literature have shown that the exposure to air pollutants increases the mortality rate caused by respiratory and cardiovascular diseases (see, for example, [33]). Particulate matter (PM) is one of the main air pollutants causing ill effects; the symbol PM_d is generally used to denote the particles having a diameter equal to or smaller than d micrometers. In most developed countries, there are regulations requiring to measure the concentrations of $PM_{2.5}$ and PM_{10}. It is evident that, at any point in time, the concentration measured for $PM_{2.5}$ at a specific site cannot be greater than the concentration of PM_{10} at the same site. As the equipment for measuring the concentration of $PM_{2.5}$ is more expensive than that for the concentration of PM_{10}, it is desirable to estimate the former from the latter. Because

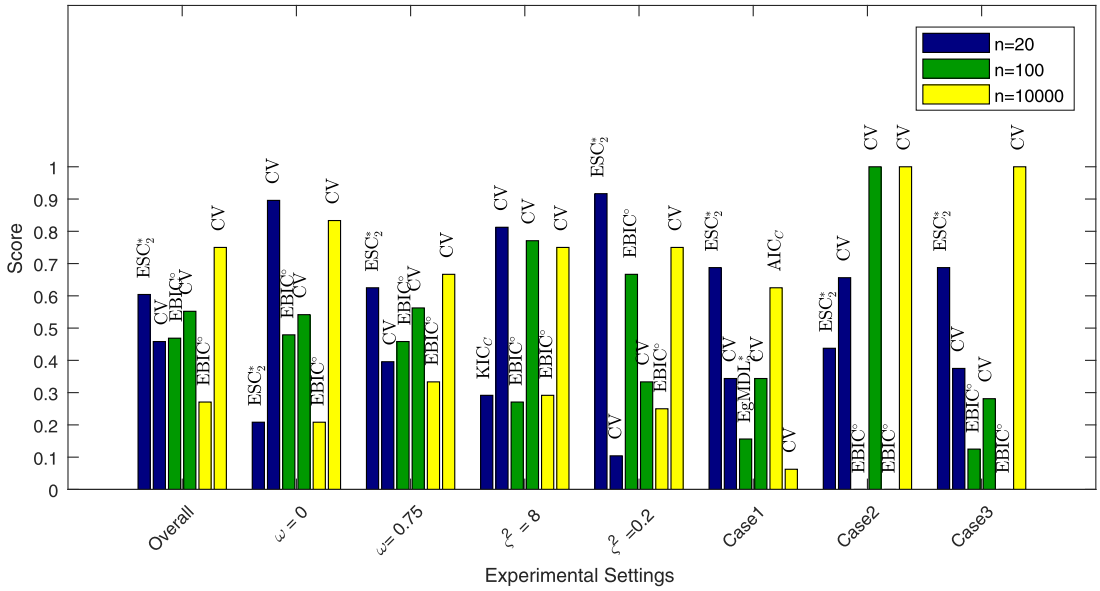

Fig. 4. Comparison of IT criteria with CV: for each scenario, the aggregated score of the best performing IT criterion is shown in the plot.

there is no clear dependence between daily measurements of the two concentrations, we treat this as a system identification problem.

5.2.2. Measurements

The New Zealand National Institute of Water and Atmospheric (NIWA) measures hourly the concentrations of $PM_{2.5}$ and PM_{10} (in $\mu g/m^3$) at various locations in Auckland (174.76°E, 36.85°S). In this study, we conduct experiments with daily measurements of PM obtained by averaging the hourly measurements. Based on the integrity and continuity of the data, we consider the measurements for the sites Patumahoe, Penrose, Takapuna, and Whangaparaoa from 30/04/2008 to 30/06/2014. The interested reader can find statistical data about these regions of Auckland in [34–37]; a map with the locations of the four sites is provided in [32]. The missing values (about 2% for each site) are imputed by applying a variant of the expectation maximization algorithm which uses smooth spline (see the function `mnimput` in [38] and the discussion in [39, Sec. 2.1.2]).

For a measurement site, let $s_1, s_2, \dots$ be the time series of log-transformed daily concentrations of $PM_{2.5}$. Similarly, $z_1, z_2, \dots$ are the log-transformed values for the concentrations of PM_{10} measured at the same site, during the same days. For an arbitrary integer $n > 0$, and for any $t \geq n$, we define the following vectors: $\underline{\mathbf{s}}(t) = [s_t \; s_{t-1} \cdots s_{t-n+1}]^\top$ and $\underline{\mathbf{z}}(t) = [z_t \; z_{t-1} \cdots z_{t-n+1}]^\top$.

5.2.3. Predictive models

We want to find a linear model which describes the relationship between the log-transformed concentration of $PM_{2.5}$ on the current day at Patumahoe and the following variables: (i) past and present log-transformed concentrations of $PM_{2.5}$ for all other three sites and (ii) past and present log-transformed concentrations of PM_{10} for the Patumahoe site. Given the particularities of the data that we analyse, we propose two different scenarios:

(i) *Full set of predictors (FullSet):* let $n = 365$. We take the response vector to be $\underline{\mathbf{s}}_{\text{PA}}(t)$ and, in order to be consistent with the previous notation, we name it $\mathbf{y}_{\text{PA}}(t)$, where PA stands for the site Patumahoe. We use the symbol $\mathbf{X}^{(i)}_{\text{PA}}(t)$ for the dictionary, in order to emphasize that it is for Scenario (i). It is given by $\mathbf{X}^{(i)}_{\text{PA}}(t) = [\underline{\mathbf{z}}_{\text{PA}}(t), \dots, \underline{\mathbf{z}}_{\text{PA}}(t-n), \underline{\mathbf{s}}_{\text{PE}}(t), \dots, \underline{\mathbf{s}}_{\text{PE}}(t-n), \underline{\mathbf{s}}_{\text{TA}}(t), \dots, \underline{\mathbf{s}}_{\text{TA}}(t-n), \underline{\mathbf{s}}_{\text{WH}}(t), \dots, \underline{\mathbf{s}}_{\text{WH}}(t-n)]$, where PE, TA, and WH represent Penrose, Takapuna and Whangaparaoa, respectively. Obviously, we have an overcomplete dictionary because the total number of predictors $p_n = 4(n+1) = 1464$ is much larger than n.

(ii) *Constrained set of predictors (ConSet):* we keep $n = 365$ and the same $\mathbf{y}_{\text{PA}}(t)$, but we reduce the total number of predictors by using empirical knowledge from air pollution scientists. More precisely, the dictionary $\mathbf{X}^{(ii)}_{\text{PA}}(t)$ has two blocks. The first block contains log-transformed measurements of PM_{10} from Patumahoe site: $\underline{\mathbf{z}}_{\text{PA}}(t), \underline{\mathbf{z}}_{\text{PA}}(t-1), \dots, \underline{\mathbf{z}}_{\text{PA}}(t-10)$, $\underline{\mathbf{z}}_{\text{PA}}(t-182), \underline{\mathbf{z}}_{\text{PA}}(t-183), \underline{\mathbf{z}}_{\text{PA}}(t-184)$, $\underline{\mathbf{z}}_{\text{PA}}(t-n+2), \underline{\mathbf{z}}_{\text{PA}}(t-n+1), \underline{\mathbf{z}}_{\text{PA}}(t-n)$, i.e., focusing on the present, the recent past, six months ago and one year ago. The second block comprises the log-transformed concentrations of $PM_{2.5}$ collected from the sites Penrose, Takapuna, and Whangaparaoa, in the same days as the measurements within the first block. Note that $p_n = 4 \cdot 17 = 68$, thus $p_n < n$.

5.2.4. Performance evaluation

In each run we use a frame of length $3n$ from the data, corresponding to three consecutive years of measurements. The first two years are used for training the linear predictor and the last year for evaluating it. For the rth run, let t_r be the last day of the second year. With the convention that $(\cdot)^{\text{sc}}$ is used to distinguish between the two prediction scenarios, $\mathbf{y}_{\text{PA}}(t_r)$ and $\mathbf{X}^{\text{sc}}_{\text{PA}}(t_r)$ are used for training. Note that the response vector and the columns of the dictionary are centred and standardised as explained in Section 5.1. The resulting vector of linear parameters is further used together with the dictionary $\mathbf{X}^{\text{sc}}_{\text{PA}}(t_r + n)$ in order to produce the estimate $\hat{\mathbf{y}}^{\text{sc}}_{\text{PA}}(t_r + n)$. This procedure is applied for $\text{N}_{TR} = 100$ runs, where the values of t_r ($r = \overline{1, 100}$) are chosen as follows. We take t_0 to be 30/04/2008 and $t_1 = t_0 + 2n - 1$; then $t_{r+1} = t_r + 8$ for $r = \overline{1, 99}$.

The experiment is repeated by selecting another site than Patumahoe to be the site for which the level of $PM_{2.5}$ is predicted; the measurements of $PM_{2.5}$ from the other three sites as well as the measurements of PM_{10} from the current site are used for building the dictionaries as it was described above. In each case, the normalized mean square error (NMSE) is computed by applying the formula:

$$\text{NMSE}^{\text{sc}}_{\text{site}} = 100 \times \frac{\sum_{r=1}^{\text{N}_{TR}} \left\| \exp\left[\mathbf{y}_{\text{site}}(t_r + n)\right] - \exp\left[\hat{\mathbf{y}}^{\text{sc}}_{\text{site}}(t_r + n)\right] \right\|^2}{\sum_{r=1}^{\text{N}_{TR}} \left\| \exp\left[\mathbf{y}_{\text{site}}(t_r + n)\right] \right\|^2}. \quad (28)$$

Table 2
Predictive models for air pollution data: the stopping rules applied to MPA are listed in the first column of the table. The values of NMSE computed with formula in (28) are shown in the other columns. In each column, the best result is represented in bold and is underlined, the results which are within a range of 1% from the best value on that column are shown in bold, and the results which are larger than 10% of the minimum value on that column are shown in italic.

	Patumahoe		Penrose		Takapuna		Whangaparaoa	
	FullSet	ConSet	FullSet	ConSet	FullSet	ConSet	FullSet	ConSet
AIC_C	*7.96*	5.52	*5.97*	**3.66**	*7.96*	**5.08**	*4.99*	**3.13**
KIC	*8.09*	5.61	*6.06*	**3.64**	*8.07*	**5.12**	*5.08*	**3.13**
KIC_C	*7.60*	5.62	*5.77*	**3.63**	*7.80*	**5.13**	*4.80*	**3.13**
BIC	*8.09*	5.82	*6.06*	**3.65**	*8.07*	5.27	*5.08*	3.16
EBIC	*8.09*	5.87	*6.06*	**3.65**	*8.07*	5.28	*5.08*	3.18
EBIC*	**6.16**	*5.97*	4.24	3.80	6.80	5.48	**3.35**	3.24
EBIC°	6.61	*6.17*	4.18	3.96	6.61	*6.08*	3.50	3.33
SC_1	*7.29*	5.82	*5.32*	**3.64**	*7.37*	5.25	*4.52*	3.16
SC_2	*7.33*	5.82	*5.34*	**3.64**	*7.38*	5.25	*4.55*	3.17
ESC_1	6.62	5.87	*4.88*	**3.64**	*7.02*	5.27	*3.95*	3.19
ESC_2	6.61	5.87	*4.86*	**3.64**	*7.02*	5.27	*3.91*	3.19
ESC_1^*	**6.16**	*5.98*	**4.03**	3.82	**6.19**	5.50	**3.35**	3.25
ESC_2^*	**6.17**	*5.98*	**4.04**	3.83	**6.19**	5.51	**3.35**	3.25
$gMDL_1$	*7.29*	5.81	*5.33*	**3.64**	*7.37*	5.24	*4.53*	3.16
$gMDL_2$	*7.33*	5.81	*5.34*	**3.64**	*7.38*	5.24	*4.55*	3.16
$EgMDL_1$	6.63	5.86	*4.88*	**3.64**	*7.03*	5.26	*3.95*	3.18
$EgMDL_2$	6.63	5.86	*4.88*	**3.64**	*7.04*	5.26	*3.93*	3.18
$EgMDL_1^*$	**6.14**	*5.98*	**4.04**	3.81	**6.19**	5.48	**3.35**	3.24
$EgMDL_2^*$	**6.16**	*5.98*	**4.04**	3.81	**6.19**	5.48	**3.35**	3.25
MMLU	*7.35*	5.82	*5.35*	**3.65**	*7.39*	5.25	*4.55*	3.16
$MMLG_1$	*7.99*	5.83	*5.86*	**3.63**	*7.80*	5.21	*4.95*	**3.16**
$MMLG_2$	*8.00*	5.83	*5.87*	**3.63**	*7.80*	5.21	*4.96*	**3.16**
CV	*7.30*	**5.42**	*4.95*	3.81	6.63	**5.13**	*4.34*	3.17

Note that the denominator corresponds to a trivial predictor having all coefficients equal to zero.

The values of NMSE computed by applying various stopping rules for MPA are outlined in Table 2. They show that it is highly recommended to employ the "extended criteria" for the FullSet dictionary ($n < p_n$), but not for the ConSet dictionary ($n > p_n$). This observation is perfectly in line with the theoretical grounds on which these criteria are based (see the discussion in Section 3). In the family of the "extended criteria", the best results are produced by ESC_{alt}^* and $EgMDL_{alt}^*$, where alt=1,2. There is no stopping rule that is clearly the best one for the ConSet dictionary: for Patumahoe the minimum NMSE is given by CV, for Penrose by $MMLG_1$, for Takapuna by AIC_C and for Whangaparaoa by KIC. This situation is so partly because almost all the criteria tend to work relatively well when the total number of predictors is constrained to be $p_n = 68$.

The results reported for simulated data and air pollution data can be reproduced by using the Matlab code available at https://www.stat.auckland.ac.nz/~cgiu216/PUBLICATIONS.htm.

6. Final remarks

In this paper, we have investigated various IT criteria that can be employed as stopping rules for MPA. As all of them depend on df given by the trace of the hat matrix, we provided some theoretical results about the hat matrix. One of the main contributions of this work is an efficient algorithm for computing df when $n \gg p_n$ (big data).

For all the IT criteria that we have analyzed, the goodness-of-fit term is essentially the same and what differentiates them is the penalty term. We consider a classification of the selection rules into the families $\{\mathcal{F}_i\}_{i=1}^5$, based on the form of the penalty. With the convention that alt $= 1, 2$, we have:

- $\mathcal{F}_1$ – the penalty term depends only on the sample size and the degrees of freedom: AIC_C, KIC, KIC_C, BIC;
- $\mathcal{F}_2$ – the penalty term depends explicitly on both the response $\mathbf{y}$ and the data $\mathbf{X}$: SC_{alt}, $gMDL_{alt}$, MMLU, MMLG;
- $\mathcal{F}_3$ – an extra penalty term, directly proportional to $\nu \ln \vartheta$, is added to some of the criteria from $\mathcal{F}_1$ and $\mathcal{F}_2$: EBIC, ESC_{alt}, $EgMDL_{alt}$;
- $\mathcal{F}_4$ – an extra penalty term, directly proportional to $\ln \vartheta^*$, is added to some of the criteria from $\mathcal{F}_1$ and $\mathcal{F}_2$: EBIC*, ESC_{alt}^*, $EgMDL_{alt}^*$;
- $\mathcal{F}_5$ – the penalty $4\mathrm{df}_m \ln p_n$ is added to BIC from $\mathcal{F}_1$: EBIC°.

For simulated data, it follows from the experiments described in [32] and in the previous sections that $\mathcal{F}_4$ produces the best results when the number of samples is smaller than the number of predictors ($n < p_n$), whereas $\mathcal{F}_5$ is the winner when $n \geq p_n$. $\mathcal{F}_1$ works well only when n is small and the SNR is high, or when n is large and the data is memoryless.

In the experiment with air pollution data where the full set of predictors is considered (FullSet), $\mathcal{F}_4$ yields the minimum NMSE for all four sites. This confirms what we have already observed from simulations: $\mathcal{F}_4$ is superior to other families of criteria when $n < p_n$. It is worth pointing out that the use of $\mathcal{F}_4$ for selecting the atoms from a large dictionary for which $p_n = 1464$ (FullSet) leads to prediction results that are close to those obtained when the dictionary is constrained to contain only $p_n = 68$ atoms (ConSet). The constrained dictionary has been built by using prior knowledge from environmental chemistry. In what concerns the case $n > p_n$ for air pollution data (ConSet), $\mathcal{F}_1$ and $\mathcal{F}_2$ demonstrate an advantage with respect to other families.

An important outcome of the comprehensive set of experiments that we have conducted with both simulated and air pollution data is: $\mathcal{F}_4$-criteria introduced in this work are the best stopping rules when $n < p_n$. Their performance is similar to and sometimes superior to CV, but the computational complexity is lower. It is not surprising that all the members of $\mathcal{F}_4$ are modified versions of "extended criteria", which have been designed for the situation when the total number of predictors is large. However, our study shows that it is important how the alteration of the original criteria is

done: even if $\mathcal{F}_3$ and $\mathcal{F}_4$ are obtained by modifying the same set of original criteria, the performance of $\mathcal{F}_3$ is worse than that of $\mathcal{F}_4$. The unique member of $\mathcal{F}_5$ is also derived from an "extended" criterion, but it is particularly useful for large sample sizes when either the SNR is low or the data have memory.

We have examined above the performance for families of IT criteria and not individually because the total number of selection rules evaluated in this study is very large.

We conclude that MPA can be successfully applied for big data ($n \gg p_n$) as well as for overcomplete dictionaries ($n < p_n$) if the stopping rule is properly chosen.

Acknowledgements

The authors are grateful to one of the reviewers for pointing out the modification that led to the expressions of the criteria in (11)–(13). Dr. C.D. Giurcăneanu would like to thank to Dr. Elizabeth Somervell from NIWA for many useful discussions on air pollution data.

Appendix A. Some properties of the hat matrix

We need the following technical results:

Result 1. Let $\bar{\mathbf{x}}, \mathbf{y} \in \mathbb{R}^n$ such that $||\bar{\mathbf{x}}|| = 1$ and $\mathbf{y} \neq \mathbf{0}$. If $\mathbf{P} = \bar{\mathbf{x}}\bar{\mathbf{x}}^\top$, we have:

$$||(\mathbf{I} - \nu\mathbf{P})\mathbf{y}|| \leq ||\mathbf{y}|| \text{ if } \nu \in (0, 1]. \tag{A.1}$$

The equality is achieved if and only if $\bar{\mathbf{x}}^\top\mathbf{y} = 0$.

Proof. The result can be established by observing that $||(\mathbf{I} - \nu\mathbf{P})\mathbf{y}||^2 = \mathbf{y}^\top(\mathbf{I} - \nu\mathbf{P})^2\mathbf{y}$ and the largest eigenvalue of the symmetric matrix $(\mathbf{I} - \nu\mathbf{P})^2$ is equal to one. Then (A.1) is a consequence of the well-known Rayleigh inequality. □

Result 2. For $\mathbf{y} \in \mathbb{R}^n$, we have that $||\mathbf{A}_m\mathbf{y}|| \leq ||\mathbf{y}||$ when $m \geq 1$.

Proof. Using the notation $\mathbf{A}_0 = \mathbf{I}$, Result 1 implies that $||(\mathbf{I} - \nu\mathbf{P}_{s(j+1)})(\mathbf{A}_j\mathbf{y})|| \leq ||\mathbf{A}_j\mathbf{y}||$ for all $0 \leq j \leq m-1$. This leads straightforwardly to Result 2. □

Result 3. The following identity holds true for $m \geq 1$:
$\text{tr}(\mathbf{B}_{m+1} - \mathbf{B}_m) = \nu\text{tr}(\mathbf{P}_{s(m+1)}\mathbf{A}_m)$.

Proof. We can readily write the identities: $\text{tr}(\mathbf{B}_{m+1} - \mathbf{B}_m) = \text{tr}[(\mathbf{I} - \mathbf{A}_{m+1}) - (\mathbf{I} - \mathbf{A}_m)] = \text{tr}(\mathbf{A}_m - \mathbf{A}_{m+1}) = \text{tr}\left[\mathbf{A}_m - (\mathbf{I} - \nu\mathbf{P}_{s(m+1)})\mathbf{A}_m\right] = \nu\text{tr}(\mathbf{P}_{s(m+1)}\mathbf{A}_m)$. □

Now we show that, at each step of the MPA, the increase of df is at most ν. This can be recast as a property of the hat matrix:

Proposition 2. *For $m \geq 1$, we have*

$$\text{tr}(\mathbf{B}_{m+1}) - \text{tr}(\mathbf{B}_m) \leq \nu. \tag{A.2}$$

The equality holds if and only if $\bar{\mathbf{x}}_{s(m+1)}^\top\bar{\mathbf{x}}_{s(j)} = 0$ for all $j \in \{1, \ldots, m\}$.

Proof. An important consequence of Result 3 is that, for proving Proposition 2, it suffices to demonstrate the inequality $\text{tr}(\mathbf{P}_{s(m+1)}\mathbf{A}_m) \leq 1$. The fact that $\text{rank}(\mathbf{P}_{s(m+1)}) = 1$ implies $\text{rank}(\mathbf{P}_{s(m+1)}\mathbf{A}_m) \leq 1$. Additionally, we have that $(\mathbf{P}_{s(m+1)}\mathbf{A}_m)\bar{\mathbf{x}}_{s(m+1)} = (\bar{\mathbf{x}}_{s(m+1)}\bar{\mathbf{x}}_{s(m+1)}^\top)\mathbf{A}_m\bar{\mathbf{x}}_{s(m+1)} = (\bar{\mathbf{x}}_{s(m+1)}^\top\mathbf{A}_m\bar{\mathbf{x}}_{s(m+1)})\bar{\mathbf{x}}_{s(m+1)}$, which demonstrates that the only non-zero eigenvalue of $\mathbf{P}_{s(m+1)}\mathbf{A}_m$ is $\bar{\mathbf{x}}_{s(m+1)}^\top\mathbf{A}_m\bar{\mathbf{x}}_{s(m+1)}$. Hence, we get:

$$|\text{tr}(\mathbf{P}_{s(m+1)}\mathbf{A}_m)| = |\bar{\mathbf{x}}_{s(m+1)}^\top\mathbf{A}_m\bar{\mathbf{x}}_{s(m+1)}| \leq ||\bar{\mathbf{x}}_{s(m+1)}||\,||\mathbf{A}_m\bar{\mathbf{x}}_{s(m+1)}|| \tag{A.3}$$

$$= ||\mathbf{A}_m\bar{\mathbf{x}}_{s(m+1)}|| \leq ||\bar{\mathbf{x}}_{s(m+1)}|| = 1. \tag{A.4}$$

The inequality in (A.3) is obtained by using the properties of the scalar product [31, Th. 1.1], while the inequality in (A.4) is based on Result 2.

The equality holds in (A.2) if and only if we have simultaneously equalities in (A.3) and (A.4). As we know from Result 1 that $||(\mathbf{I} - \nu\mathbf{P}_{s(j)})\bar{\mathbf{x}}_{s(m+1)}|| \leq ||\bar{\mathbf{x}}_{s(m+1)}||$ for any $j \in \{1, \ldots, m\}$, the only possibility for having equality in (A.2) is $\bar{\mathbf{x}}_{s(m+1)} \in \bigcap_{j=1}^m \text{Ker}(\mathbf{P}_{s(j)})$. The condition is equivalent to $\bar{\mathbf{x}}_{s(m+1)}^\top\bar{\mathbf{x}}_{s(j)} = 0$ for all $j \in \{1, \ldots, m\}$. □

Remark 1. For all $m \geq 1$, one can show that $\text{tr}(\mathbf{B}_{m+1}) - \text{tr}(\mathbf{B}_m) \geq -\nu$ by using Result 3 and inequality (A.4). In practice, it is observed that $\text{tr}(\mathbf{B}_{m+1}) - \text{tr}(\mathbf{B}_m)$ can be negative, hence is not guaranteed that df increases at each iteration of MPA.

In general, $\mathbf{B}_m$ is not a projection matrix. As a square matrix is a projector if and only if it is idempotent (see, for example, [31, Th. 2.1]), we check when $\mathbf{B}_m^2 = \mathbf{B}_m$.

Proposition 3. *(i) If $\nu \in (0, 1)$, then $\mathbf{B}_m$ is not idempotent for all $m \geq 1$.*

(ii) Consider the following conditions: (c_1) $m \geq 2$; (c_2) $\nu = 1$;
(c_3) $\bar{\mathbf{x}}_{s(i)}^\top\bar{\mathbf{x}}_{s(j)} = 0$ for all $i, j \in \{1, \ldots, m\}$ with property $i > j$. If all these conditions are satisfied, then $\mathbf{B}_m$ is idempotent and symmetric.

Proof. (i) Using the identity in (2), it is easy to show that $\mathbf{B}_m$ is idempotent if and only if $\mathbf{A}_m$ is idempotent. Another important observation is that $\det(\mathbf{A}_m) = (1 - \nu)^m$, where $\det(\cdot)$ denotes the determinant. This is a consequence of the fact that $\det(\mathbf{I} - \nu\mathbf{P}_{s(j)}) = 1 - \nu$ for $j \in \{1, \ldots, m\}$. As $\nu \in (0, 1)$, we have $\det(\mathbf{A}_m) \in (0, 1)$. Therefore, $\mathbf{A}_m$ is not idempotent because the determinant of an idempotent matrix can only be zero or one.

(ii) We have from hypothesis that $\mathbf{P}_{s(i)}\mathbf{P}_{s(j)} = \mathbf{0}$ for $m \geq i > j \geq 1$. This property together with the identities in (4) and (5) lead to the conclusion that $\mathbf{A}_m = \mathbf{I} - (\mathbf{P}_{s(m)} + \cdots + \mathbf{P}_{s(1)})$. It follows from (2) that $\mathbf{B}_m = \mathbf{P}_{s(m)} + \cdots + \mathbf{P}_{s(1)}$. It is easy to check that $\mathbf{B}_m$ is idempotent and symmetric. □

Remark 2. The second part of Prop. 3 can be understood in connection with the result from [31, p. 44] which says that a sufficient condition for $\mathbf{B}_m = \sum_{k=1}^m (-1)^{k+1} \sum_{m \geq j_k > j_{k-1} > \cdots > j_1 \geq 1} \mathbf{P}_{s(j_k)}\mathbf{P}_{s(j_{k-1})} \cdots \mathbf{P}_{s(j_1)}$ (see (2)-(3)) to be the orthogonal projector onto $\text{Sp}\left(\bar{\mathbf{x}}_{s(1)}, \ldots, \bar{\mathbf{x}}_{s(m)}\right)$ is:

$$\mathbf{P}_{s(i)}\mathbf{P}_{s(j)} = \mathbf{P}_{s(j)}\mathbf{P}_{s(i)} \text{ for all } i, j \in \{1, \ldots, m\}. \tag{A.5}$$

Remark 3. In order for the strong condition (c_3) in Proposition 3 to be fulfilled, we need $m \leq n$.

At the end of this analysis, we prove the following result:

Proposition 4. *(i) If $\nu \in (0, 1)$, then $\mathbf{A}_m^\top\mathbf{A}_m + \mathbf{B}_m^\top\mathbf{B}_m \neq \mathbf{I}$ for all $m \geq 1$.*

(ii) If the conditions (c_1)–(c_3) from Proposition 3(ii) are satisfied, then $\mathbf{A}_m^\top\mathbf{A}_m + \mathbf{B}_m^\top\mathbf{B}_m = \mathbf{I}$.

Proof. (i) Assume that

$$\mathbf{A}_m^\top\mathbf{A}_m + \mathbf{B}_m^\top\mathbf{B}_m = \mathbf{I}, \text{ or equivalently,} \tag{A.6}$$

$$2\mathbf{A}_m^\top\mathbf{A}_m - \mathbf{A}_m - \mathbf{A}_m^\top = \mathbf{0}. \tag{A.7}$$

Let $\mathbf{v}$ be an eigenvector of $\mathbf{A}_m$ corresponding to the eigenvalue λ. Using the fact that $\mathbf{A}_m\mathbf{v} = \lambda\mathbf{v}$ together with (A.7), we get: (a) $\mathbf{A}_m^\top\mathbf{v} = \frac{\lambda}{2\lambda-1}\mathbf{v}$, which shows that $\frac{\lambda}{2\lambda-1}$ is an eigenvalue for $\mathbf{A}_m^\top$. As the eigenvalues of $\mathbf{A}_m^\top$ are the same with the eigenvalues of $\mathbf{A}_m$, it follows that $\frac{\lambda}{2\lambda-1}$ is also an eigenvalue for $\mathbf{A}_m$. (b) $(\mathbf{A}_m^\top\mathbf{A}_m)\mathbf{v} = \frac{\lambda^2}{2\lambda-1}\mathbf{v}$, which demonstrates that $\frac{\lambda^2}{2\lambda-1}$ is an eigenvalue for $\mathbf{A}_m^\top\mathbf{A}_m$. Since we know from the proof of Proposition 3(i) that $\det(\mathbf{A}_m) =$

$(1-\nu)^m > 0$, we have that the symmetric matrix $\mathbf{A}_m^\top \mathbf{A}_m$ is positive definite. Hence, all the entries of $\mathbf{v}$ are real-valued and $\lambda > 1/2$.

The considerations above imply that the positive numbers λ and $\frac{\lambda}{2\lambda-1}$ are eigenvalues for $\mathbf{A}_m$. An important consequence of Result 2 is that both eigenvalues are less than or equal to one. However, if $\lambda \leq 1$, then $\frac{\lambda}{2\lambda-1} \geq 1$. Therefore, we need to have $\lambda = 1$. In other words, all eigenvalues of $\mathbf{A}_m$ are equal to one, which we know it is not possible because $\det(\mathbf{A}_m) = (1-\nu)^m < 1$. We obtained this contradiction because we have assumed that the identity in (A.6) holds true.

(ii) We have from the proof of Proposition 3(ii) that $\mathbf{A}_m$ is idempotent and symmetric, therefore the identity in (A.7) is true. □

Remark 4. At the mth step of MPA, we obtain the estimate $\hat{\mathbf{y}}_m = \mathbf{B}_m \mathbf{y}$ and the error $\mathbf{e}_m = \mathbf{y} - \hat{\mathbf{y}}_m = \mathbf{A}_m \mathbf{y}$. In general, $||\hat{\mathbf{y}}_m||^2 + ||\mathbf{e}_m||^2 \neq ||\mathbf{y}||^2$.

Appendix B. Proofs for the results presented in Section 4

Proof of Theorem 1: We rewrite the expression of $\mathbf{A}_m$ by expanding (4) and (5) and grouping the terms which have in common the first and the last predictor. The terms that contain a single predictor are written separately. Hence, we get:

$$\mathbf{A}_m = \mathbf{I} - \nu \sum_{j=1}^{m} \bar{\mathbf{x}}_{s(j)} \bar{\mathbf{x}}_{s(j)}^\top + \sum_{a=1}^{p_n} \sum_{b=1}^{p_n} \sum_{i} (-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a \bar{\mathbf{x}}_a^\top) \Xi_{i,ab} (\bar{\mathbf{x}}_b \bar{\mathbf{x}}_b^\top). \tag{B.1}$$

The number of terms in the sum over i and the values of the exponents $q_{i,ab}$ are not important, as they are not computed explicitly. We do not need to calculate the factors $\Xi_{i,ab}$, but it is helpful to write down the following chain of identities:

$$\begin{aligned}
&\mathrm{tr}\Big[(-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a \bar{\mathbf{x}}_a^\top) \Xi_{i,ab} (\bar{\mathbf{x}}_b \bar{\mathbf{x}}_b^\top)\Big] \\
&= (-\nu)^{q_{i,ab}} \mathrm{tr}\Big[\bar{\mathbf{x}}_a (\bar{\mathbf{x}}_a^\top \Xi_{i,ab} \bar{\mathbf{x}}_b) \bar{\mathbf{x}}_b^\top\Big] \\
&= (-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a^\top \Xi_{i,ab} \bar{\mathbf{x}}_b) \mathrm{tr}\big(\bar{\mathbf{x}}_a \bar{\mathbf{x}}_b^\top\big) \\
&= (-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a^\top \Xi_{i,ab} \bar{\mathbf{x}}_b) d_{ba},
\end{aligned} \tag{B.2}$$

where d_{ba} denotes the entry of $\mathbf{D}$ located in the bth row and the ath column.

For writing the equations more compactly, we define the matrix $\mathbf{G}_m$ whose entries are:

$$g_{m,ab} = \sum_{i} (-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a^\top \Xi_{i,ab} \bar{\mathbf{x}}_b) \text{ for } 1 \leq a, b \leq p_n. \tag{B.3}$$

Even if not explicitly expressed, the quantities in the right-hand side depend on m because they are affected by the predictor selected at the mth step. It follows from Eqs. (B.1)–(B.3) that

$$\begin{aligned}
\mathrm{tr}(\mathbf{A}_m) &= n - m\nu + \sum_{a=1}^{p_n} \sum_{b=1}^{p_n} \sum_{i} \mathrm{tr}\Big[(-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a \bar{\mathbf{x}}_a^\top) \Xi_{i,ab} (\bar{\mathbf{x}}_b \bar{\mathbf{x}}_b^\top)\Big] \\
&= n - m\nu + \sum_{a=1}^{p_n} \sum_{b=1}^{p_n} \sum_{i} (-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a^\top \Xi_{i,ab} \bar{\mathbf{x}}_b) d_{ba} \\
&= n - m\nu + \sum_{a=1}^{p_n} \sum_{b=1}^{p_n} g_{m,ab} d_{ba} \\
&= n - m\nu + \mathrm{tr}(\mathbf{G}_m \mathbf{D}).
\end{aligned} \tag{B.4}$$

The recurrence relation in (16) leads to

$$\begin{aligned}
&\mathrm{tr}(\mathbf{A}_{m+1}) - \mathrm{tr}(\mathbf{A}_m) \\
&= -\nu \mathrm{tr}\big(\bar{\mathbf{x}}_{s(m+1)} \bar{\mathbf{x}}_{s(m+1)}^\top \mathbf{A}_m\big) \\
&= -\nu \bar{\mathbf{x}}_{s(m+1)}^\top \mathbf{A}_m \bar{\mathbf{x}}_{s(m+1)} \\
&= -\nu + \nu^2 \sum_{j=1}^{m} d_{s(m+1)s(j)}^2 \qquad [\text{see } (B.1)]
\end{aligned}$$

$$-\nu \sum_{a=1}^{p_n} \sum_{b=1}^{p_n} \bar{\mathbf{x}}_{s(m+1)}^\top \bar{\mathbf{x}}_a \left[\sum_{i} (-\nu)^{q_{i,ab}} (\bar{\mathbf{x}}_a^\top \Xi_{i,ab} \bar{\mathbf{x}}_b)\right] \bar{\mathbf{x}}_b^\top \bar{\mathbf{x}}_{s(m+1)} \tag{B.5}$$

$$= -\nu + \nu^2 \sum_{j=1}^{m} d_{s(m+1)s(j)}^2 \qquad [\text{see } (B.3)]$$

$$-\nu \sum_{a=1}^{p_n} \sum_{b=1}^{p_n} d_{s(m+1)a} g_{m,ab} d_{bs(m+1)} \tag{B.6}$$

$$\begin{aligned}
&= -\nu \\
&\quad + \sum_{b=1}^{p_n} \left[\nu^2 d_{s(m+1)b} \sum_{j=1}^{m} \mathbb{1}_{[s(j)=b]} - \nu \sum_{a=1}^{p_n} d_{s(m+1)a} g_{m,ab}\right] d_{bs(m+1)},
\end{aligned} \tag{B.7}$$

where $\mathbb{1}_{[s(j)=b]}$ is equal to one if $s(j) = b$ and zero otherwise. Note that $\sum_{j=1}^{m} \mathbb{1}_{[s(j)=b]}$ represents how many times the predictor $\bar{\mathbf{x}}_b$ was selected in the first m steps of MPA. The identities in (B.4) and (B.7) imply that

$$\begin{aligned}
\mathrm{tr}(\mathbf{A}_{m+1}) &= n - (m+1)\nu + \sum_{b=1}^{p_n} \sum_{a=1}^{p_n} g_{m,ab} d_{ba} \\
&\quad + \sum_{b=1}^{p_n} \left[\nu^2 d_{s(m+1)b} \sum_{j=1}^{m} \mathbb{1}_{[s(j)=b]} - \nu \sum_{a=1}^{p_n} d_{s(m+1)a} g_{m,ab}\right] \\
&\quad \times d_{bs(m+1)}.
\end{aligned} \tag{B.8}$$

Additionally, we have from (B.4) that

$$\mathrm{tr}(\mathbf{A}_{m+1}) = n - (m+1)\nu + \sum_{b=1}^{p_n} \sum_{a=1}^{p_n} g_{m+1,ab} d_{ba}. \tag{B.9}$$

Comparing the expressions of $\mathrm{tr}(\mathbf{A}_{m+1})$ given in the equations above, we conclude that the only entries of $\mathbf{G}_{m+1}$ that are different from those of $\mathbf{G}_m$ are located in the $s(m+1)$th row. Moreover, the following recursive formula holds for these entries:

$$\begin{aligned}
g_{m+1,s(m+1)b} &= g_{m,s(m+1)b} \\
&\quad + \nu^2 d_{s(m+1)b} \sum_{j=1}^{m} \mathbb{1}_{[s(j)=b]} - \nu \sum_{a=1}^{p_n} d_{s(m+1)a} g_{m,ab}, \\
&\quad \text{for } 1 \leq b \leq p_n.
\end{aligned} \tag{B.10}$$

The vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^{p_n}$ that appear in Theorem 1 are defined as follows. The bth entry of $\mathbf{w}$, w_b, is a count for how many times the predictor $\bar{\mathbf{x}}_b$ was selected in the first m steps of MPA. Note that (19) is a straightforward consequence of this definition. The bth entry of $\mathbf{v}$ is

$$v_b = \nu^2 d_{s(m+1)b} w_b - \nu \sum_{a=1}^{p_n} d_{s(m+1)a} g_{m,ab}, \text{ for } 1 \leq b \leq p_n, \tag{B.11}$$

which leads to (17). Additionally, Eqs. (B.10) and (B.11) prove the recursion in (18). The formula for the computation of the degrees of freedom which is given in (20) follows immediately from (B.8)–(B.11).

Proof of Lemma 1: (i) The use of (17) for evaluating the vector $\mathbf{v}$ at the $\ell(j+1)$th iteration leads to the following calculations:

$$\begin{aligned}
\mathbf{v}_{\ell(j+1)}^\top &= -\nu \mathbf{d}^\top \mathbf{G}_{\ell(j+1)-1} + \nu^2 \mathbf{d}^\top \odot \mathbf{w}_{\ell(j+1)-1}^\top \\
&= -\nu [1 \ \tilde{\mathbf{d}}^\top] \begin{bmatrix} \mathbf{g}_{\ell(j)}^\top \\ \tilde{\mathbf{G}}_{\ell(j+1)-1} \end{bmatrix} + \nu^2 (j + j_0) \mathbf{h}^\top + \nu^2 [0 \ \tilde{\mathbf{d}}^\top \odot \tilde{\mathbf{w}}_{\ell(j+1)-1}] \\
&= -\nu \mathbf{g}_{\ell(j)}^\top + \nu^2 (j + j_0) \mathbf{h}^\top + \mathbf{r}_{\ell(j+1)}^\top.
\end{aligned}$$

By employing (18) and the identity above, we get (21). Furthermore, we have that $\mathbf{v}_{\ell(j+1)}^\top = \mathbf{g}_{\ell(j+1)}^\top - \mathbf{g}_{\ell(j)}^\top = (1 -$

$\nu)\left(\mathbf{g}_{\ell(j)}^{\top}-\mathbf{g}_{\ell(j-1)}^{\top}\right)+\nu^2[j+j_0-(j-1+j_0)]\mathbf{h}^{\top}+\mathbf{r}_{\ell(j+1)}^{\top}-\mathbf{r}_{\ell(j)}^{\top}$, which proves the identity in (22).

(ii) The identity in (23) is a straightforward consequence of (22).

Proof of Prop. 1: We need two technical results:

Result 4. Let $(a_\mu)_{\mu\geq 1}$ be a sequence of real numbers. We define $S_\mu=\sum_{j=1}^{\mu}a_j(1-\nu)^{\mu-j}$, where $\mu\geq 1$ and $\nu\in(0,\ 1)$. If $\lim_{\mu\to\infty}a_\mu=0$, then $\lim_{\mu\to\infty}S_\mu=0$.

Proof. Let $\epsilon>0$. Because $\lim_{\mu\to\infty}a_\mu=0$, there is $\mu_0\geq 1$ such that $|a_\mu|<(\epsilon\nu)/2$ for all $\mu\geq\mu_0$. Similarly, because $\lim_{\mu\to\infty}(1-\nu)^\mu=0$, there is $\mu_1\geq 1$ such that $(1-\nu)^\mu<\frac{\epsilon/2}{\sum_{j=1}^{\mu_0}|a_j|}$ for all $\mu\geq\mu_1$. For all $\mu\geq(\mu_0+\mu_1)$, we have: $|S_\mu|=\left|\sum_{j=1}^{\mu_0}a_j(1-\nu)^{\mu-j}+\sum_{j=\mu_0+1}^{\mu}a_j(1-\nu)^{\mu-j}\right|<(1-\nu)^{\mu-\mu_0}\sum_{j=1}^{\mu_0}|a_j|+\frac{\epsilon\nu}{2}\sum_{j=\mu_0+1}^{\mu}(1-\nu)^{\mu-j}<\frac{\epsilon}{2}+\frac{\epsilon\nu}{2}\frac{1}{\nu}=\epsilon$. This shows that $\lim_{\mu\to\infty}S_\mu=0$. □

Result 5. Let $(a_\mu)_{\mu\geq 1}$ be the same as in Result 4. This time we define $S_\mu=\frac{1}{\mu}\sum_{j=1}^{\mu}a_j$ for $\mu\geq 1$. We show that $\lim_{\mu\to\infty}a_\mu=0$ implies $\lim_{\mu\to\infty}S_\mu=0$.

Proof. Let ϵ and μ_0 have the same significance like in the proof of Result 4. As $\lim_{\mu\to\infty}\frac{1}{\mu}=0$, there is $\mu_1\geq 1$ such that $\frac{1}{\mu}<\frac{\epsilon/2}{\sum_{j=1}^{\mu_0}|a_j|}$ for all $\mu\geq\mu_1$. Hence, for all $\mu\geq(\mu_0+\mu_1)$, we have: $|S_\mu|<\frac{1}{\mu}\sum_{j=1}^{\mu_0}|a_j|+\frac{\epsilon\nu}{2}\frac{\mu-\mu_0}{\mu}<\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon$, which proves that $\lim_{\mu\to\infty}S_\mu=0$. □

It follows from (23) that $\lim_{\mu\to\infty}\mathbf{v}_{\ell(\mu)}^{\top}=\nu\mathbf{h}^{\top}+\lim_{\mu\to\infty}\sum_{j=2}^{\mu}(\mathbf{r}_{\ell(j)}^{\top}-\mathbf{r}_{\ell(j-1)}^{\top})(1-\nu)^{\mu-j}$. It can be easily shown that the limit in the right-hand side is $\mathbf{0}$ by applying Result 4 to each sequence $(a_\mu)_{\mu>1}$ that corresponds to an entry of the difference $(\mathbf{r}_{\ell(j)}^{\top}-\mathbf{r}_{\ell(j-1)}^{\top})_{j>1}$. This proves the result in (24).

Similarly, Result 5 implies that $\lim_{\mu\to\infty}\frac{1}{\mu}\sum_{j=2}^{\mu}(\mathbf{r}_{\ell(j)}^{\top}-\mathbf{r}_{\ell(j-1)}^{\top})=\mathbf{0}$. Hence, we have $\lim_{\mu\to\infty}\frac{\mathbf{r}_{\ell(\mu)}^{\top}}{\mu}=\mathbf{0}$. Additionally, Eq. (21) leads to $\lim_{\mu\to\infty}\frac{\mathbf{g}_{\ell(\mu)}^{\top}}{\mu}=-\frac{1}{\nu}\lim_{\mu\to\infty}\frac{\mathbf{v}_{\ell(\mu)}^{\top}}{\mu}+\nu\mathbf{h}^{\top}\lim_{\mu\to\infty}\frac{\mu+j_0}{\mu}+\frac{1}{\nu}\lim_{\mu\to\infty}\frac{\mathbf{r}_{\ell(\mu)}^{\top}}{\mu}$. This identity together with the previous results show that (25) is true.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.sigpro.2018.09.033

References

[1] F. Ghido, I. Tabus, Sparse modeling for lossless audio compression, IEEE Trans. Audio Speech Lang. Process. 21 (2013) 14–28.
[2] R. Shumway, D. Stoffer, Time Series Analysis and its Applications. With R Examples, 3rd ed., Springer Science+Business Media, 2011.
[3] P. Bühlmann, S. van de Geer, Statistics for High-dimensional Data. Methods, Theory and Applications, Springer-Verlag, 2011.
[4] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, IEEE Trans. Signal Process. 41 (1993) 3397–3415.
[5] P. Bühlmann, T. Hothorn, Boosting algorithms: regularization, prediction and model fitting, Stat. Sci. 22 (2007) 477–505.
[6] A. Barron, A. Cohen, W. Dahmen, R. DeVore, Approximation and learning by greedy algorithms, Ann. Stat. 36 (2008) 64–94.
[7] A. Sancetta, Greedy algorithms for prediction, Bernoulli 22 (2016) 1227–1277.
[8] F. Li, C. Triggs, B. Dumitrescu, C. Giurcăneanu, On the number of iterations for the matching pursuit algorithm, in: Proceedings of the Twenty-fifth European Signal Processing Conference (Eusipco 2017), Kos, Greece, 2017, pp. 181–185.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. data Mining, Inference, and Prediction, 2, Springer Science+Business Media, 2008.
[10] T. Hastie, Comment: boosting algorithms: regularization, prediction and model fitting, Stat. Sci. 22 (2007) 513–515.
[11] R. Tibshirani, Regression analysis and selection via the lasso, J. R. Stat. Soc. Ser. B 58 (1996) 267–288.
[12] P. Bühlmann, Boosting for high-dimensional linear models, Ann. Stat. (2006) 559–583.
[13] C. Stein, Estimation of the mean of a multivariate normal distribution, Ann. Stat. (1981) 1135–1151.
[14] J. Ye, On measuring and correcting the effects of data mining and model selection, J. Am. Stat. Assoc. (1998) 120–131.
[15] B. Efron, The estimation of prediction error: covariance penalties and crossvalidation, J. Am. Stat. Assoc. (2004) 619–632.
[16] J. Chen, Z. Chen, Extended Bayesian information criteria for model with large model spaces, Biometrika 95 (2008) 759–771.
[17] T. Roos, P. Myllymäki, J. Rissanen, MDL denoising revisited, IEEE Trans. Signal Process. 57 (2009) 3347–3360.
[18] D. Schmidt, E. Makalic, The consistency of MDL for linear regression models with increasing signal-to-noise ratio, IEEE Trans. Signal Process. 60 (2012) 1508–1510.
[19] R. Foygel, M. Drton, Extended Bayesian information criteria for Gaussian graphical models, in: J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, A. Culotta (Eds.), Advances in Neural Information Processing Systems, 23, 2010, pp. 604–612.
[20] A. Owrang, M. Jansson, A model selection criterion for high-dimensional linear regression, IEEE Trans. Signal Process. 66 (2018) 3436–3446.
[21] C. Giurcăneanu, S. Razavi, A. Liski, Variable selection in linear regression: several approaches based on normalized maximum likelihood, Signal Process. 91 (2011) 1671–1692.
[22] E. Artin, The Gamma Function, Holt, Rinehart and Winston, Inc., 1964.
[23] C. Hurvich, C.L. Tsai, Regression and time series model selection in small samples, Biometrika 76 (1989) 297–307.
[24] C. Hurvich, J. Simonoff, C.L. Tsai, Smoothing parameter selection in nonparametric regression using an improved akaike information criterion, J. R. Stat. Soc.: Ser. B Part 2 (1998) 271–293.
[25] J. Cavanaugh, A large-sample model selection criterion based on Kullback's symmetric divergence, Stat. Probab. Lett. 42 (1999) 333–343.
[26] A.K. Seghouane, M. Bekara, A small sample model selection criterion based on Kullback's symmetric divergence, IEEE Trans. Signal Process. 52 (2004) 3314–3323.
[27] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (1978) 461–464.
[28] J. Rissanen, Information and Complexity in Statistical Modeling, Springer Verlag, 2007.
[29] M. Hansen, B. Yu, Model selection and the principle of minimum description length, J. Am. Stat. Assoc. 96 (2001) 746–774.
[30] D. Schmidt, E. Makalic, MML invariant linear regression, in: Proceedings of the Twenty-second Australasian Joint Conference on Artificial Intelligence, 2009, pp. 312–321.
[31] H. Yanai, K. Takeuchi, Y. Takane, Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition, Springer Science+Business Media, 2011.
[32] F. Li, C. Triggs, B. Dumitrescu, C. Giurcăneanu, Supplemental material to: the matching pursuit algorithm revisited: a variant for big data and new stopping rules (2018), doi:10.1016/j.sigpro.2018.09.033.
[33] B. Brunekreef, S.T. Holgate, Air pollution and health, Lancet 360 (2002) 1233–1242.
[34] Statistics New Zealand, 2013 Census, Quickstats about Patumahoe, 2013a, http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-about-a-place.aspx?request_value=13452&tabname=Work&sc_device=pdf. Technical Report
[35] Statistics New Zealand, 2013 Census, Quickstats about Penrose, 2013b, http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-about-a-place.aspx?request_value=13427&parent_id=13171&tabname=&sc_device=pdf. Technical Report
[36] Statistics New Zealand, 2013 Census, Quickstats about Devonport–Takapuna Local Board Area, 2013c, http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-about-a-place.aspx?request_value=13613&tabname=&sc_device=pdf. Technical Report
[37] Statistics New Zealand, 2013 Census, Quickstats about Hibiscus and Bays Local Board Area, 2013d, http://www.archive.stats.govt.nz/Census/2013-census/profile-and-summary-reports/quickstats-about-a-place.aspx?request_value=13610&parent_id=13170&tabname=&sc_device=pdf. Technical Report
[38] W. Junger, A.P. de Leon, Multivariate Time Series Data Imputation, 2012, R-package `mtsdi` documentation, https://www.cran.r-project.org/web/packages/mtsdi/mtsdi.pdf.
[39] Y. Yang, Predictive Modelling for the Concentration of air Pollutant $PM_{2.5}$, 2016, (Master's thesis), University of Auckland